Recommendations

of the

Job Training Longitudinal Survey Research

Advisory Panel

to

Office of Strategic Planning and Policy Development

Employment and Training Administration

U.S. Department of Labor

The Panel:

Howard Bloom
Robert Boruch
Michael Borus
Judith Gueron
Alan Gustman
Peter Rossi
Fritz Scheuren
Marshall Smith
Frank Stafford
Ernst Stromsdorfer - Chair

November 1985

HD 5715.2 .U53 1985

TABLE OF CONTENTS

Section		Page
A.	The Problem	1
B.	What Have We Learned from the Present Set of Studies of CETA?	5
C.	Analysis Problems That Affect the Validity and Reliability of Program Evaluation	12
D.	The Recommendations of the Panel	20
E.	Recommendations for DOL Management of the JTPA Evaluation	28
F.	Conclusions	31

Appendix I - Optimal Sample Sizes for Evaluating JTPA

Appendix II - Review of Findings and Suggestions for Future CLMS Research on the Net Impact of CETA

Appendix III - Evaluating Manpower Training: Some Notes on Data Handling Issues

Appendix IV - A Summary Comparison of the Job Training Longitudinal Survey and the Continuous Longitudinal Manpower Survey

Appendix V - Issues to Consider in Establishing Classical Experiments of Employment and Training Programs

A. The Problem

To conform with the legal requirements of the Job Training Partnership Act (JTPA), the United States Department of Labor (DOL) must "...evaluate the effectiveness of programs authorized under this act...," specifically with respect to "...[the effect of the Act on] increases in employment and earnings for participants, reduced income support costs, increased tax revenues, duration in training and employment situations, information on the postenrollment labor market experience of program participants for at least a year following their termination from such programs, and comparable information on other employees or trainees of participating employers." (Stress added. See JTPA, Section 454.(a).)

The DOL has initiated an evaluation of the JTPA by means of the Job Training Longitudinal Survey (JTLS) with its accompanying planned comparison group sample, the Survey of History of Work (SHOW). Field data collection for JTLS has begun. These data are complemented by Quick-Turnaround (QT) data sampled from the management information systems of the Service Delivery Areas (SDAs). The QT data have been collected since October 1983.

The general design of JTLS is derived from the earlier Continuous Longitudinal Manpower Survey (CLMS) that has been used to evaluate the Comprehensive Employment and Training Act (CETA) in conjunction with the Current Population Survey (CPS) and social security earnings records linked to both data sets. (See Appendix IV for an outline of the JTLS design and its comparison with the CLMS. See also Westat, 1984 (September).)

The statistical methods and analysis design used to analyze the labor market impacts of the Comprehensive Employment

and Training Act represented a reasonable scientific approach at the time of its initiation in 1974, especially given the legal and resource constraints then faced by DOL. Subsequent experience with the CLMS/CPS evaluation of CETA, as well as with advances in statistical and econometric technique, have shown that the estimates of net impact of CETA are not reliable and that the true net impacts of CETA are still open to question. Since the methods intended to be used to evaluate the JTPA are broadly the same as those used to evaluate CETA, there is considerable likelihood that the validity of the net impact evaluation of JTPA will also be subject to unresolvable doubt if the existing analysis design remains unchanged.

The Department of Labor is most serious in its desire to avoid this possible outcome. To improve its chances of adopting the most valid and reliable methods to evaluate the JTPA, the DOL has appointed the JTLS Technical Advisory Panel (the Panel) to assist it in resolving several critical problems that have been revealed by these earlier CETA evaluations.

The Panel has been charged to review the major evaluations of CETA that employ the CLMS/CPS data set in order to assess the general validity of these results, taken as a whole. In addition, the Panel is to recommend alternative methods to deal with the unresolved methodological problems that have severely qualified the CLMS/CPS evaluations of CETA--problems that will affect the evaluation of the JTPA unless corrected.

A Fundamental Issue in Program Evaluation. The CLMS presents information on economic outcomes for a sample of individuals who enrolled in the different subsidized training, employment and job placement programs offered under CETA. There are no observations indicating what the labor market outcomes would have been had these same people not gone through the

program. Yet to measure the impact of these programs, it would be ideal to compare, for each individual, outcomes that would be observed had the individual attended the programs with the outcomes that would have been observed had the same individual never enrolled in the program. In the absence of such information, analysts have tried to find individuals who never enrolled in the program but who are sufficiently like those who did so that the outcomes of these comparison individuals are a valid approximation of those outcomes that would have been observed for the program participants had they not entered the program. The labor market or other experiences of those who attended the program are compared, after statistical adjustment of any measurable differences, to that of the individuals who were not participants, and the difference is attributed as a net effect of the program. This is called a quasi-experimental analysis design.

An alternataive strategy for determining which individuals are sufficiently similar so that their outcomes can be compared is to conduct an experiment in which program-eligible individuals are assigned randomly to the program. If this random assignment is not compromised by either the eligible individuals or the program managers, program participants will not differ statistically in any systematic or unmeasured way from those who do not enroll in the program. (More precisely, there will be differences, but the expected value of those statistical differences will be zero.) Such experiments require the withholding of services from some program-eligible individuals on a random basis. This is a difficult proposition to implement, but, if successful, this random assignment will yield valid estimates of net program impact.

As indicated above, an alternative to a classic experimental design is a quasi-experimental design where people

who are enrolled in the program are compared with individuals who did not enroll but had similar economic, educational and demographic characteristics. Unfortunately, there is a problem with this approach. Those who enter employment and training programs may do so for a variety of economic reasons and non-economic reasons. However, the decision to enter a program is a result of systematic differences between those who enroll and those who do not, even if both groups have the same observable demographic characteristics and economic histories before enrollment. For example, in the case of women, those who enter the program may have more desire to work in future years than those who do not, and this may mean that in the postprogram period, program completers are likely to choose a job that allows them to work more hours or has a different career potential than those who chose not to enroll. The result will be an overestimation of the impact of the program.

This problem has been called the problem of selection or, selection bias. Econometricians have been working to solve the problem of selection mostly by trying to better understand the program enrollment decision, which in turn will help provide an understanding of how those legally-eligible persons who are and are not enrolled in the program differ from one another. Once this understanding is gained, it should be possible to correct statistically for these differences. It is difficult to tell how long it will be until such efforts bear fruit. Success is unlikely to be achieved in the short time needed to produce policy relevant analyses that would allow improvements to be legislated in the JTPA program. Therefore, as discussed more fully below, the Panel has decided to recommend the experimental option at this time, an option that is the most likely one to generate answers that are usable for evaluating the JTPA and for designing improvements in it. However, it is also prudent to continue development work on the econometric front; use of the

quasi-experimental analysis design in the long run may be the only way we can regularly measure nationally representative program outcomes and be able to do this without having to deny program treatments to some.

A more detailed discussion of the different kinds of selection bias, of the weaknesses of the quasi-experimental approach taken in evaluating CETA, and of the Panel's recommendations now follows.

Plan of the Report. The next section describes the net impact results that have been estimated for CETA and assesses the degree of internal validity of these results--how accurately the true program effects are measured--for policy guidance. Section C then discusses the key methodological problems faced by the CETA CLMS/CPS design and quasi-experimental models like it, such as the JTLS/SHOW design. This discussion sets the stage for the Panel recommendations on JTPA evaluation design in Section D and related implementation recommendations in Section E. Finally, the analysis underlying the Panel recommendations is backed up in part by five analytical appendices.

B. What Have We Learned from the Present Set of Studies of CETA?

A substantial portion of the CLMS/CPS data is currently available as a public use data set. Over the past several years the DOL, to its credit, has commissioned a number of competing analyses and evaluations of the CETA, all using the CLMS/CPS data set. Other researchers not sponsored by DOL have also analyzed these data and contributed to the technical debate. (See Ashenfelter and Card, 1984; Bassi, 1983 and 1984; Bassi, et al., 1984; Bloom and McLaughlin, 1982; Dickinson, Johnson and West,

1984; Finifter, 1983; Fraker and Maynard, 1985; Heckman and Robb, 1983; Geraci, 1984; Lalonde, 1984; and Westat, 1984 (April).)

With the exception of the analysis by Dickinson, Johnson and West (1985), these analyses of CETA are roughly consistent in their estimation of the net labor market impact of CETA. Tables 1 and 2 detail the estimated effects of CETA on adult men and women. (More detail on this summary analysis is presented in Appendix II. Separate results for youths are not reported because it is not clear how best to measure labor market impacts on youth and because youth have little past labor market experience to serve as a control for the problem of selection bias that is discussed below.)

In examining Tables 1 and 2 note that the findings reported represent only one set among several (alternative) net impact estimates generated by each study. An attempt was made to select estimates that were most representative of the overall findings and conclusions of each study or which represented the most recent, and presumably, most reliable estimates. However, these results do not reflect all of the nuances and qualifications addressed by each study. Nor do they show the wide variety of results that may be obtained when seemingly small changes are made in the sample composition or in the statistical techniques of analysis. Indeed, Ashenfelter and Card (1984) specifically demonstrate that the results one gains depends on the explicit choice of econometric methods and that it is hard to choose among these methods as to which is most correct.

Thus, the several authors made a variety of adjustments to the CLMS/CPS data base and each used somewhat different econometric models to arrive at their results. Nevertheless, the tables shown reflect the general pattern of results obtained by

Table 1. Net impact by CETA on annual earnings for adult women (rounded to the nearest 100 in 1978 dollars: CLMS data set)

Study	All Women	White Women	Non-White Women
Bassi (1983) Table 3.			
Classroom Training	NR	200	600*
On-the-job Training	NR	400	1,500*
Work Experience	NR	-120	1,000*
Public Service Employment	NR	700*	800*
Bloom and McLaughlin (1984) Tables 5 and 10.			
Overall	700 to 1,100**	700 to 1,100**	500 to 800**
Classroom Training	700 to 1,200**	700 to 1,100**	600 to 900**
On-the-job Training	600 to 900**	600 to 1,000**	400 to 700*
Work Experience	700 to 1,100**	700 to 1,200**	400 to 700**
Dickinson, et al. (1984) Table IV-3.			
Overall	0	150	-200
Classroom Training	0	NR	NR
On-the-job Training	0	NR	NR
Work Experience	-500*	NR	NR
Public Service Employment	500	NR	NR
Direct Referral	200	NR	NR
Finifter[1] (1983) Table X.			
1977 Findings	500	NR	NR
1978 Findings	1,200	NR	NR
Geraci (1984) Table 1.			
Classroom Training	1,200**	NR	NR
On-the-job Training	900**	NR	NR
Work Experience	300	NR	NR
Public Service Employment	1,100**	NR	NR
Westat (April 1984) Tables 5-1 and 5-3.			
FY 1976 Cohort			
1977 Impact	NR	400**	300**
1978 Impact	NR	700**	800**
FY 1977 Cohort			
1978	NR	500**	800**
Ashenfelter and Card (1984) Table 5.[2]	600 to 1,400**	NR	NR

Source: Appendix II.

Notes: [1]Finifter did not report the statistical significance of his summary net impact estimates.

[2]Original estimates in 1967 dollars were inflated to 1978 dollars.

* and ** indicate statistical significance at nominal 0.05 and 0.01 levels, respectively.

NR means the results were not reported or estimated.

Table 2. Net impact by CETA on annual earnings for adult men (rounded to the nearest 100 in 1978 dollars: CLMS data set)

Study	All Men	White Men	Non-White Men
Bassi (1983) Table 3.			
Classroom Training	NR	NR	800
On-the-job Training	NR	NR	2,100*
Work Experience	NR	NR	-300
Public Service Employment	NR	NR	0*
Bloom and McLaughlin (1984) Tables 5 and 10.			
Overall	200	-100	500*
Classroom Training	200	200	200
On-the-job Training	200	-200	1,200
Work Experience	-100	-200	200
Dickinson, et al. (1984) Table IV-2.			
Overall	-700**	-800	-500
Classroom Training	-300	NR	NR
On-the-job Training	-400	NR	NR
Work Experience	-500	NR	NR
Public Service Employment	-800**	NR	NR
Direct Referral	-700*	NR	NR
Finifter[1] (1983) Table X.			
1977 Findings	-100	NR	NR
1978 Findings	-1,500	NR	NR
Geraci (1984) Table 1.			
Classroom Training	400	NR	NR
On-the-job Training	600*	NR	NR
Work Experience	-600*	NR	NR
Public Service Employment	-200	NR	NR
Westat (April 1984) Tables 5-1 and 5-3.			
FY 1976 Cohort			
1977 Impact	NR	0	-100
1978 Impact	NR	0	-300
FY 1977 Cohort			
1978	NR	500**	700**
Ashenfelter and Card (1984) Table 4.	-400 to 300	NR	NR

Source: Appendix II.

Notes: [1]Finifter did not report the statistical significance of his summary net impact estimates.

* and ** indicate statistical significance at nominal 0.05 and 0.01 levels, respectively.

NR means the results were not reported or estimated.

each study and thereby provide an approximate basis for comparing them.

General Results: Adult Women. The set of studies generally agree that the returns in annual earnings increases to adult women who participated in CETA were positive, large, and nominally statistically significant. For all women, the estimate of net program impact ranges from $300 to $1,200 per year, depending on whether one views the impact of work experience or the classroom training treatments. The impacts of on-the-job training and public service employment fall within this range. It is possible that white women fare better than non-white women for any program treatment, but the reasons for this are not known other than the obvious one of possible race discrimination. In particular, it is not known whether the quality and quantity of any given treatment varies between the two sex-race groups. Finally, if these are the true net impacts and if they do not decay over four or five years, the CETA program appears to be cost-effective. For instance, given cost outlays of approximately $3,700 per participant-year in 1984 (while a participant enrolls on average for half-year or less), the payback periods are often less than two years, which implies a very high rate of return on investment to classroom training.

General Results: Adult Men. The net impact of CETA on annual earnings for adult men is very small or zero, and may even be negative, based on these results. None of the studies yields consistent results; negative and positive effects reside side by side and there is no particular pattern of statistical significance. While one could make a case for cost-effective results for adult women, this is not possible for adult men. At best we must be agnositc about the program's net effect on adult men, but, as noted, the net impact of CETA could even be negative, depending on what one believes motivated adult men to

join CETA and/or what econometric method one chose to correct for this motivation and related behavior. (This motivation issue encompasses in part the problem of selection bias--the main issue confronting the valid analysis of JTPA's net impact).

Assessment. Given the methodological problems with the CLMS and CPS data sets, there is currently little consensus on the choice of methods to estimate net program impacts using these particular data. The general pattern of findings for CETA is consistent with patterns found for its predecessor, The Manpower Development and Training Act--at least for adult women. (See Ashenfelter, 1978.) However, the range of results within and across studies is disturbingly large and no particular point estimate can be said to be the correct one. The results by themselves cannot serve as a reliable guide to education and training investment decisions for CETA or its successor, JTPA. In particular, if, as is likely the case, selection bias exists, the returns to women may be overstated, since many of the women who were CETA participants may have entered the labor market in the absence of CETA. Additionally, they may also have acquired training and educational services on their own.

In contrast, the negative results reported for men could exist for a variety of reasons. First, adult men could be simply using the program as an income maintenance substitute. That is, some proportion of adult males may have joined the program just to gain access to the monthly income maintenance payments. They may have had no serious intent to treat CETA as a method to enhance their labor market prospects. In other words, there was no program treatment as such. It is reasonable to suggest this as a possible behavior since the Panel does not believe that such a large mass of individuals would knowingly injure themselves and there is no obvious reason why a simple treatment like job search assistance would reduce earnings. (See

Dickinson, et al. 1985.) This possibility has never been appropriately investigated.

Next, it could be the case that employers may stigmatize adult males who enter subsidized employment and training programs as somehow being less productive and, therefore, employers may be less likely to hire them.

Finally, it is also possible that the CPS data set did not provide the appropriate comparison groups, the variables to standardize for differences between participants and the comparison groups, or both. The work by Fraker and Maynard (1985) shows that each of these conditions is a possibility. Fraker and Maynard were able to approximate the true experimental results of the National Supported Work Demonstration for AFDC women using a comparison group generated from the CPS, but were not able to achieve the same result for youths. The CPS data set contained both insufficient numbers of the appropriate low-income and disadvantaged youths as well as incorrect variables to describe the complex and often volatile economic experiences of youths.

Given the existing CLMS/CPS data, econometric methods may not be able to unravel these conundrums, though a reanalysis and synthesis of the existing set of studies has been funded by DOL. Indeed, the more sophisticated methods require a minimum set of appropriate data in order to successfully estimate program effects. In this light, a well-designed classical experiment is likely to shed light on these issues in a more cost-effective way and in a shorter time frame than would be the case if the JTPA evaluation were based on a quasi-experimental design. However, as we shall see below, practical problems exist with the classical experimental approach as well.

In the meantime, the Panel offers the judgment on the currently available evidence that CETA may be a reasonable social investment for disadvantaged adult women. Its value to adult men and to youths of both sexes is at best very unclear. We do not know.

C. Analysis Problems That Affect the Validity and Reliability of Program Evaluations.

This section discusses two broad analysis issues that affect the validity and reliability of the evaluation of social programs such as CETA and JTPA. The recommendations of the Panel are based on the implications of this discussion. The two issues are:

- Selection bias: What is it and what can be done about it?
- Sample design: What is the appropriate sampling procedure for an analytic evaluation?

Issues of data cleaning, dropping "invalid" observations and definition of program treatment based on duration of program attendance that are specific to the pecularities of the CLMS/CPS data set are not discussed. To be sure, these are important issues, but are of somewhat less importance and generality, given the focus of the Panel on the JTLS. (Specific work on these issues has been recently funded by DOL and is being executed by Dr. James Heckman and his associates and these issues are also discussed in Appendices II and III.)

Selection Bias. Selection bias is the most serious and important problem that faces a quasi-experimental evaluation design such as the CETA CLMS/CPS design or the recently initiated JTLS/SHOW design. Failure to overcome selection bias can lead to

unknown over- or underestimation of net program impacts. Econometric efforts to adjust for this bias can lead to the wide range of results we see in Tables 1 and 2. State-of-the-art statistical techniques simply can not overcome certain data or design problems.

Indeed, the most significant work to date on this issue is that of Robert Lalonde (1984). Again, using the data from the National Supported Work Experiment as did Fraker and Maynard, Lalonde asks the question: "What will happen if non-experimental econometric methods are used to estimate net program impacts?" There are two basic results: first, methods that pass sensible econometric specification tests perform no better than those that don't; second, many perfectly reasonable methods lead to answers that are all over the map and reasonable persons will disagree as to which answers to choose as "correct."

As noted in Section A above, selection bias occurs when program eligible individuals select themselves in or out of the program in a non-random, purposive manner or when program managers do the same to equally eligible (in a legal sense) individuals. All social programs properly have selection incorporated into their legal and administrative structure. This type of selection can be modelled statistically to the extent that it is formally specified in the law or regulations and variables exist to measure directly the selection criteria. For instance, selection into a given program may be limited to individuals who are receiving AFDC. But selection behavior is often also based on informal criteria, some of which may be unmeasurable, such as the motivation or work attitudes of the program-eligible individual or informal criteria used by program managers to select "more deserving" or "moresuitable" individuals. (When "more suitable" individuals are selected, this is called "creaming".)

Selection has at least four components that can affect the validity of a net program impact estimate. In a quasi-experimental evaluation design, each has to be adjusted for statistically:

1. A legally eligible individual decides to participate or not participate in JTPA.

2. A program manager chooses one legally eligible individual over another--e.g., two individuals are "economically disadvantaged", but one is judged "more disadvantaged" than another based on the value system of the program manager.

3. A program manager, having admitted an individual to a program, then directs that individual to one type of treatment rather than another, perhaps because it is judged the individual will benefit more in one treatment than another or perhaps just because a slot is available.

4. Having enrolled in JTPA, the participant decides to accept one type of treatment or treatment mix rather than another.

Given a quasi-experimental design such as the CETA CLMS/CPS design, it is imperative to statistically adjust each of these types of selection if one is to estimate statistically unbiased measures of program effect by type of treatment. However, as a practical matter it is proving to be extremely difficult to model (in technical terms, statistically identify) each of these four types of selection. In point of fact, none of the CETA evaluations discussed above claims to have properly modelled, and therefore, statistically controlled for, all aspects of selection bias in the CLMS/CPS data set. Indeed, the necessary variables to do this--such as a measure of the motivation to work--simply are not in the data set. The available variables in the CPS data set are very limited. And, most of the critical variables that one would use to model

selection, such as motivation, are of the class of unmeasurable variables. This is true even for the potentially more useful SHOW and JTLS data bases.

The major methodological differences in the various evaluations cited in this report all hinge on efforts to adjust for selection. At their core the differences between Westat (1984, April) and Dickinson, et al. (1985) seen to arise because the alignment of longitudinal data, the extent to which the participant and comparison groups have similar attachments to the labor force and defining the program treatment based on length of stay in the program. Selection is also the issue with respect to controlling for the pre-program dip in earnings that many participants have experienced and that can affect the amount and direction of net program impacts if not adjusted for correctly.

It is the problem of selection that has driven analysts toward the use of more sophisticated econometric modelling as well as toward the use of relatively expensive longitudinal data. It is important to note the conclusions from the work of Heckman and Robb (1983)--neither sophisticated econometric techniques nor longitudinal data are necessarily solutions to the problem of selection bias. However, the jury on the efficacy of these econometric approaches simply is not in; hence, as the Panel recommends below, additional work on these issues should be performed.

In the meantime, the most potentially fruitful solution to the problem of selection in any evaluation is the use of a classical randomized experiment. Here one has the greatest assurance that the program participants and the control group of program eligible individuals are on the average alike in every important sense except that one group has had the program treatment while the otherwise "identical" group has not. (In

particular, this helps resolve the problem of selection bias introduced by unmeasurable variables).

Again, the Panel emphasizes the critical requirement of better addressing the problem of selection before significant time and resources are devoted to the collection and analysis of new non-experimental data sets like JTLS/SHOW. Without a resolution of the selection problem, an evaluation will not conform to the charge of the Act: to relate the outcomes of participants to..."comparable information on other employees or trainees of participating employers." Again, it is important to emphasize that selection bias renders groups noncomparable along important behavioral dimensions that affect measured program outcomes. And, selection bias will always exist in the evaluation of programs when using a quasi-experimental design such as the JTLS/SHOW.

Sample Design and Size. The existing sample design and sample sizes of the JTLS are intended for a description of the JTPA program rather than an analytic investigation of the net impact of JTPA. (See Appendix III for the technical details.) Also, the sample sizes among target and treatment groups may not have been specified appropriately. (See Appendix I for further details.)

The current JTLS sample design is set forth in "Summary Description of JTLS Study Design" (Westat, September, 1984). A review of this design--as it is contrasted with CLMS in Appendix IV--suggests that it will recreate several of the sampling and methodological problems of the CLMS design.

Sample survey methods and experimental design techniques are related. The two approaches can be integrated into a single data collection effort. (See Sedransk, 1967.) Two

distinct viewpoints could be followed in examining CLMS and JTLS design issues. Deming (1950) looks at survey designs as being separable into two types: "enumerative" and "analytic." Broadly speaking, enumerative surveys are designed to produce good estimates of population means and totals. Analytic surveys are developed to focus on making inferences about which of two or more alternatives is to be preferred. Variance estimation is more important in analytic surveys than in enumerative surveys. As a result, the optimal estimation of population totals sometimes has to take second place.

The JTLS Design. The sampling design for JTLS is a strictly enumerative type. To obtain optimal estimates of totals the usual approach in a multi-stage stratified probability sample is to construct strata that are as homogeneous as possible and then to select one primary sampling unit per stratum. The approach taken for the JTLS was to group the SDA's into strata and then select one SDA per stratum. "Each of these strata was composed of SDAs with similar social and economic characteristics (e.g., percent unemployed, percent urban, percent below the poverty level, etc.)." In other words, the strata were made as homogeneous as possible on background covariates and no attempt was made to look at the extent to which SDA training modalities in treatment design and delivery differed. Yet, any evaluation of JTPA should focus, among other things, on the effects of treatment design and delivery.

However, the current JTLS design is an entirely appropriate one if the focus is on learning about SDA enrollee characteristics at entry. Given that little was known about how this new program would operate, it is hard to find fault with a design strategy that took as its starting point the measurement of general background information. It was however a mistake, given the analytic focus of JTLS, to have taken only one SDA per

stratum. First, by doing this, variance estimation problems arise. Second, stratification factors and treatment effects tend to be confounded by this design. It should be noted that the earlier CLMS design avoided these problems, to some degree, by selecting two primary sampling units per stratum. Finally, different sample sizes may result for given target/treatment groups depending on the sampling strategy used.

Sample Sizes. Corollary to changing the sample design is a recommended change in the sampling decision rule used to specify sample sizes for any given group, such as adult males taking classroom trianing.

The estimates in Table 3 demonstrate the differences in cell sizes one arrives at relative to the current JTLS design if a sampling approach is used based on a sampling decision rule that takes into consideration the relative policy importance of different program treatments based on their relative per participant costs: classroom training, on-the-job training and job placement services. (Technically speaking, a modified Bayesian-decision theoretic model is being employed.)

A major outcome of this reanalysis of the sample size issue is that it is likely to be distributed much differently among target groups and treatments than the current JTLS design has indicated. And, it appears that the overall sample size required could be much larger in total than that suggested by the current JTLS design. This implies that, given the budget constraint that the DOL is under, all 24 target group and treatment combinations in Table 3 could not be analyzed either under a quasi-experimental design or a classical random design. Choices will have to be made as to which target group and treatment combinations will be analyzed and the sequence in which they will be analyzed.

Table 3. Sample size[1] estimates for an experimental design to evaluate JTPA using the modified Bayesian-Decision Theoretic Model

Target Group	Classroom Training[2]		On-the-job Training[2]		Job Placement Service[2]	
	Number	Cost in Thousands	Number	Cost in Thousands	Number	Cost in Thousands
Male						
White						
Young (16-25)	1,200	900	1,200	900	1,100	825
Old (26-71)	1,600	1,200	1,400	1,050	1,100	825
Black						
Young (16-25)	1,200	900	600	450	1,300	975
Old (26-71)	1,400	1,050	800	600	1,400	1,050
Female						
White						
Young (16-25)	1,500	1,125	800	600	1,000	750
Old (26-71)	1,800	1,350	1,000	750	1,000	750
Black						
Young (16-25)	1,400	1,050	500	375	1,000	750
Old (26-71)	1,600	1,200	500	375	1,100	825
TOTAL:	11,700	8,775	6,800	5,100	9,000	6,750
TOTAL, based on the current JTLS Design:[3]	5,000	---	5,000	---	2,000	---

Source: Appendix I and Westat, 1984 (September).

Notes: [1]Divided equally between comparison group observations and participant group observations.

[2]The assumption is that there are three waves of interviews per individual observation that sum to a total cost of $750. The data are to be collected over a 4-year postprogram period. Thus, for data collection alone, the assumption is that it would cost $900,000 to collect the data necessary to evaluate the impact of classroom training for young adult males. Data collection costs to evaluate the six older male/treatment combinations would be about $5,775,000 over the projected 4-year data collection period (7,700 x $750).

[3]Although the SHOW sample size was never decided upon, the assumption here is that the JTLS participant sample and the SHOW comparison group sample would have equal sample sizes, i.e., for classroom training, JTLS equals 2,500, and therefore, so will SHOW.

As a working estimate, the Panel judged that about $750 would be needed to collect and electronically code three waves of data per program participant or control group member. If this judgment is accurate, about 21 million dollars would be required to collect the necessary data to evaluate the 24 target/treatment groups in Table 3. This expenditure would be spread out over a four to five year period. However, funds would also be required to conduct the analysis of program impact based on these data. Funds would also be required for the recommended process analysis, tracking individuals over time to minimize sample attrition, the analysis of program-eligible individuals in each experimental SDA site who are not members of the experiment and for managing the day-to-day logistics of the experiment itself. (See the Panel Recommendations below.)

Finally, note that the age groups are broken into those aged 25 and older and those under age 25. This is because labor market behavior differs markedly at age 25 and older relative to under 25, especially for males. If the age cut-off were to be made at age 21 and over and under 21, these sample cell sizes would increase for the older group since the variation in behavior would be greater. This factor would obviously affect the cost of analyzing any given target/treatment group.

D. The Recommendations of the Panel

The recommendations of the Panel are strongly conditioned by the judgment that it will not be possible, to solve the problem of selection bias within the context of a quasi-experimental design such as the JTLS/SHOW; at least, not in a short enough time frame to meet Congress' needs for valid information to guide policy. Heckman and Robb (1985) demonstrate

convincingly that selection bias in principle can be adjusted using econometric techniques; however this requires the statistical identification of the model of the selection process. Even though many authors studying employment and training programs have recognized this problem, no such study using a quasi-experimental design can be said to have controlled adequately for selection bias. The Panel does not intend to set forth a counsel of despair. Rather, it is concerned that the past evaluations of CETA have consumed, and the contemplated evaluations of JTPA will consume, millions of dollars and much elapsed time. It would be extremely unfortunate if the analysis of the JTLS/SHOW design would yield the same ambiguous conclusions as has the analysis of the CLMS/CPS data base for CETA. When all is said and done, the Panel does not judge that the net impact analysis of JTPA based on the JTLS/SHOW design will resolve the debate over whether JTPA may be a good social investment. The issue of selection bias is likely to remain unsolved. It is now time to resolve this issue--at least for key target group and treatment combinations.

The Panel wishes to be both precise and modest in what it recommends. Contrary to the initial charge of the DOL for the Panel to provide several graduated options--the Panel wrestled with this approach and abandoned it--only one major option is provided. The salient points of this option are listed in Table 4. Finally, in making these recommendations, the Panel recognizes that Congress has imposed on DOL the requirement of annual assessments of JTPA that will reflect the national scope of the Act's performance and impacts. (See JTPA, Sections 169(d); 452; and 453, for instance.)

Table 4. Summary of Key Panel Recommendations

1. DOL should perform a limited number of classical experiments.

   - Focus initially on adult males and females taking classroom training, on-the-job training, and job search assistance; focus on youths next.
   - For any target/treatment group such as adult males in classroom training, operate a given experiment in two or more SDAs.
   - The experiments should reflect the operational mix and design of program treatments that exist in the field.

2. Perform process analyses in the experimental SDA sites to focus, among other things, on understanding the process of selection into the program.

3. Estimate all legally eligible individuals in each SDA experimental site. For those who do not volunteer for the experiment, draw a statistical sample and measure their propensity to enroll in JTPA or substitute programs and treatments as well as the net impact of these program treatments on labor market outcomes. This may help analyze the phenomenon of selection into programs and program treatments.

4. Maintain and improve the accuracy and timeliness of the Quick-Turnaround MIS-based data. For the foreseeable short-term future perform any national analysis on JTPA by means of linking SSA earnings records with the QT data set. Experiment with possible data sets such as the Continuous Work History Sample of the Survey of Income and Program Participation to serve as a comparison group.

5. Place on hold the JTLS/SHOW effort until experimental results are developed that provide an understanding of and methods for adjustment for selection.

6. Based on the results of the recently commissioned work of Heckman, et al., conduct further analysis of the CLMS data with the major intent to shed light on the process of selection.

Primary Recommendations

1. The DOL should perform a selected set of classical experiments over the next several years that involve random assignment of program-eligible individuals to the treatment (experimental) group and to the non-treatment (control) group. This is the key recommendation of the Panel. The intent is to use these experiments to:

   - Evaluate the net impact of JTPA for selected target/treatment groups in a set of SDAs that volunteer to participate.

   - Further, it is intended to use these experimental results and the understanding of the selection process gained thereby to improve the effectiveness of quasi-experimental designs as a strategy for program evaluation. This approach should enable the implementation of some type of JTLS/SHOW design at a later date.

This recommendation to conduct classical experiments further specifies that:

   a. The experiments be focused on key target group and treatment combinations such as adult males and females taking classroom training or job search assistance. Indeed, adult males and adult females, grouped by race or ethnic origin, are the initial groups on which to focus. Three ambiguities that are displayed in Tables 1 and 2 above must be resolved as soon as possible. Youth are the next priority candidates since JTPA devotes such a large share of its funds to aiding them. Three <u>basic</u> treatments should be considered: classroom training, on-the-job training, and job search assistance. There may be insufficient funding in the DOL budget to execute experiments on all 24 target-treatment groups shown in Table 3 and do, as well, all the other analyses and management that is also necessary. Thus, policy priorities as specified by DOL will have to be used to make choices among these groups. Some cells may have to be ignored during

the next four years. Some may be collapsed together.

b. The experiments should reflect the relative mix and quality of treatments as evidenced by the national scope of the operating program. They should not reflect any synthetic mix or combination of treatments not found in the SDA sites since the available resources will not permit this strategy in addition to an evaluation of JTPA in its on-going operational state.

c. The experiments should encompass two or more SDA locations for any given target group/treatment combination.

d. A 4-year post-program follow-up should be employed, with at least two post-program follow-up waves of questionnaires. For example, one wave can occur at 12 months after program exit to provide an early assessment of net program impact. Then there can be at least one more 36 months later. These can be a combination of telephone questionnaires and personal, face-to-face interviews in order to minimize cost.

In addition, resources must be spent to reduce the problem of crossovers from the control to the experimental treatment group. Crossovers complicate the problem of estimating unbiased program effects. In general, the existence of crossovers will result in any underestimation of net program effects if the true impact of JTPA is positive. The more favorable labor market experiences of the crossovers will raise the average post-program performance of the control group, thus lowering the net performance of the experimental group when the two groups are compared statistically.

Next, substantial resources will have to be expended to track people during the four year post-program period in order to reduce the loss due to sample attrition. Of course, this post-program follow-up period can be extended considerably

and at a very low cost by merging SSA earnings data to the field collected data set. Such a merge should be performed.

2. There should be a process analysis conducted at the experimental SDA sites to study the design, implementation and operation of JTPA SDA sites. A key focus in this analysis should be on the nature of selection of participants into the program and into its treatments by eligible individuals themselves and by program operators and managers. In addition to some of the experimental sites, a mix of well-managed SDA sites could be studied along with sites having problems in implementation and operation, funds permitting.

3. As a critical adjunct to understanding the nature of selection bias and how to adjust for it, the following important complementary analysis should be performed in each experimental SDA site. First enumerate all legally eligible individuals in the labor market area encompassed by each experimental SDA site. For those who do not volunteer for the experiment, track a sample of them with appropriate field data collection techniques, using questionnaires that are identical with those of the volunteer experimental and control groups but in addition contain any necessary behavioral questions to track potential participation in the JTPA and the reasons for doing so. If the residual population of eligibles is large relative to the research budget, sample the population appropriately. This additional data set represents at this time the best chance of analyzing and subsequently adjusting for selection bias in otherwise well-designed quasi-experimental program evaluations.

   However, one possible side effect of this procedure may be to induce more crossovers than would normally occur were every program-eligible person in a site included in the experiment. Thus, resources will have to be expended in management of the experiments to maintain the separation of the experimental individuals from the non-experimental individuals in each SDA site. To the extent that this additional crossover problem does occur, econometric techniques will have to be employed to attempt to adjust for it.

Additional Recommendations

1. DOL should maintain and improve the accuracy and timeliness of the existing Quick-Turnaround data base. In addition to the existing analysis of this data set and while the experiments are running, this sample could be canvassed 12 months after program exit (to conform with JTPA Section 454.6(b)) with a relatively simple, short mail questionnaire. This would permit DOL to measure the gross program impacts of the JTPA--impacts not statistically adjusted by contrast with a comparison group. Labor market outcomes should include those listed in Section 454.(1) of the Act, plus other outcome measures and descriptive variables, funds permitting. Again, at a later time, funds permitting, SSA longitudinal earnings data could be linked with the QT data. Data sets such as the Continuous Work History Sample (CWHS) or the Survey of Income and Program Participation (SIPP) could also be linked with the SSA longitudinal earnings data. These data sets could be used as source of comparison groups in a quasi-experimental design. However, variables that indicate program participation in JTPA do not exist in the CWHS and these variables may also be eliminated from SIPP, depending on funding decisions. This problem alone could result in a underestimation of net impact of JTPA, assuming that the true impact of JTPA is positive. (Individuals in the comparison group who participated in JTPA would have higher earnings than other comparison group members. This would raise the average performance of the comparison group and thus reduce the estimated net impact of JTPA. This may not be a serious problem since only about 4 percent to 8 percent of the program-eligible population get served by JTPA.)

2. The JTLS/SHOW sample should be placed on hold. The Panel does not judge that the issue of selection bias can be currently resolved in the context of a quasi-experimental design given the current practical problems at this time of convincingly modeling the complex selection process described above.

3. The DOL has commissioned an additional evaluation of the CETA CLMS/CPS data set that is designed to analyze and help resolve some of the differences

in findings across the recent evaluations of CETA. (See Tables 1 and 2 above for the list of studies.) Once the results of this analysis by James Heckman and his associates have taken shape, additional analysis of the CLMS data set is recommended. The express purpose of this analysis should be to further investigate the nature of the selection bias problem and other data handling issues and suggest ways to better handle them. As discussed above, the work by Fraker and Maynard (1985) suggests that some devotion of DOL funds to this effort will be useful.

Some Qualifications. The Panel does not wish to overpromise on the issue of the use of classical experiments as a resolution of DOL's historical problems in measuring net program impacts. As Appendix V points out, there are strengths and weaknesses to both the classical experimental design and the quasi-experimental design. However, properly designed and implemented, the classical experimental approach can help eliminate the problem of selection bias. Basically, the problem of attrition due to deaths, refusals and disappearances will remain but can be minimized by applying resources to tracking respondents. As noted above, the problem of crossovers will also persist.

The issue of crossovers from the control to the experimental sample group should be addressed by employing vigorous attempts to achieve voluntary agreements to the random assignment procedure and to continually prevail upon individuals and program managers to honor these agreements for the duration of the required follow-up period. It is worth devoting significant management funds to maintain the quality of the experiment. To the extent that this effort is successful, no further action need be taken. The resulting (hopefully small) number of crossovers can be analyzed both within the full set of experimental observations. As noted above, if JTPA has a

positive impact, the presence of crossovers will result in some downward bias in the estimate of net program impact.

E. Recommendations for DOL Management of the JTPA Evaluation.

Up to this point, the concern of the Panel has focused basically on the design of the JTPA evaluation. There are two more concerns to be meet in order to assure a successful evaluation--one that extends over multiple year.

First, the design has to be implemented and managed on a daily basis to assure conceptually valid and statistically accurate data for analysis.

Second, the analysis must be conducted, critiqued and revised accordingly, and in a timely fashion.

Implementation. The Panel recommends that the Department of Labor commission a qualified firm or institution--a consortium is also possible--to implement and manage the classical experimental design since a variety of design details must be reviewed, selected, implemented and monitored to ensure the success of the experiment.

The option at present is only set forth in skeletal form in this document. Extensive refining and attention to detail is required to properly flesh out the design and put it in motion. While DOL can oversee the broad conceptual design and management issues, it does not have sufficient staff resources to attend to the detailed day-to-day development, nurturing and trouble shooting required to successfully execute such a design in a way that ensures the data gathered for analysis is

conceptually valid and measures accurately what it is intended to measure.

Several issues come quickly to mind. For instance, given the obvious bureaucratic, not to mention overt policy pressures that will always exist, who will attempt to maintain the random assignment structure of such an experiment and how will this be achieved? Other practical problems can be listed, such as assisting DOL in selecting the target groups to be analyzed, encouraging SDAs to volunteer for the experiment, designing questionnaires for both the experimental volunteers and the non-volunteers that are to be tracked.

Analysis. Analysis of data requires different skills compared to the design and collection of those data, though the individuals responsible for each task must closely coordinate each of their efforts. In any case, a separate firm, institution or consortium should be commissioned to execute the analysis. Its initial research approach should be subject to critical outside review before actual analysis begins. Each analytical product should be subject to external review as the product is generated. Program analysis as complicated at the general design discussed above has a way of changing over time. Those responsible for the analysis should be allowed to respond to this evolutionary process, consistent with answering the main impact questions imposed by Congress on DOL in the JTPA. In any case, the work must be reviewed in progress by an independent party with no ax to grind. Indeed, by commissioning the set of recent papers on CETA/CLMS, the Department of Labor has clearly demonstrated its will and intent to provide program managers and Congress with as accurate and reliable program informtion as possible. (The Panel recommendation only confirms that precedent.)

Time Frame for Action. The following broad time frame is suggested:

- The design and implementation of a set of experiments for key target/treatment groups should be instituted immediately. This includes the accompanying process analysis in each experimental SDA site as well as the concomitant analysis of program participation and labor market outcomes of program-eligible individuals who do not choose to volunteer for the experiment in any SDA site.

- The QT data are already operational. DOL should continue to improve their timeliness and quality. Only if the resources--both DOL staff and funds--are available, should DOL immediately begin to set up the QT data for analysis of the national scope of JTPA. Design and implementation of the experiments, their process analysis and the concomitant tracking of legally eligible individuals who are not volunteers to the experiment have highest priority.

- The RFQ-RFP process for designing and executing the selected group of experiments should begin immediately. The various experiments can be distributed over time, with the most policy important experiments being done first. DOL will have to establish these policy priorities. The first set of experiments should be fielded within the next 12 calendar months after the RFP procedures have been completed.

F. Conclusions

The conclusions of this report can be listed briefly:

1. Assessment of the Extant CETA Evaluations:

   a. The methods used in the various analyses generally reflect the state of the art but were unable to overcome the serious problem of selection bias.

   b. Given this fact of selection bias, it is probably still the case that CETA was an effective social investment for adult women. We cannot judge the impact for adult men.

   c. No particular point estimate on net impact can be given credence for men or women. The Panel does believe the impact for women is positive and may be large enough to cover social program costs. But, beyond this, not much can be asserted.

2. Recommendations on Evaluation Design

   a. Conduct a set of classical random experiments on key target groups and treatment combinations. Among the high priority target groups are first adult men and women and then youths. Among the key treatments are classroom training in its several contexts, on-the-job training, and job search assistance.

      The issue of selection bias with respect to adult males and females should be resolved. Next, the net impact of JTPA on youth, sorted by ethnic origin and sex should be resolved for classroom training and job search assistance.

   b. Legally eligible individuals in the experimental SDAs who do not choose to volunteer for the experiment should be tracked as part of the overall study in order to gain an understanding of the selection problem.

c. There should be a process analysis of implementation and management problems that focuses on SDAs that are experimental sites. The process analysis should investigate the process of selection in particular.

d. Retain the QT. Refine it. Link it to a SSA earnings data and experiment with comparison groups drawn from data bases such as the CWHS, funds permitting.

e. Place JTLS/SHOW on hold until the experimental work helps solve or reduce the problem of selection bias.

f. Building on the findings of Heckman and his associates, further analysis of the CLMS data should be conducted with the express purpose of analyzing the problem of selection and other data handling issues. There is still much to be learned about CETA that could help in later quasi-experimental designs.

3. Recommendations on Executing the Design and Analysis —

a. An RFQ and then an RFP should be issued as soon as possible to design the set of classical random experiments. This will provide the most effective triage of qualified institutions or firms. It will assure the best level of quality for a given cost. It should also be the least time consuming approach.

b. The research process should be monitored by a firm or group with no ax to grind in order to assure adherence to DOL's policy needs and maintain a quality level that will inspire public confidence in the final research products.

SELECT BIBLIOGRAPHY

Ashenfelter, Orley. "Estimating the Effect of Training on Earnings." The Review of Economics and Statistics. February 1978.

Ashenfelter, Orley and David Card. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." Working Paper No. 1489. Cambridge, MA: National Bureau of Economic Research. November 1984.

Bassi, Laurie J. "The Effect of CETA on the Postprogram Earnings of Participants." Journal of Human Resources. Fall 1983.

Bassi, Laurie J. "Estimating the Effect of Training Programs with Non-Random Selection. The Review of Economics and Statistics. February 1984.

Bassi, Laurie J., et al. Measuring the Effect of CETA on Youth and the Economically Disadvantaged. Washington, D.C.: The Urban Institute. April 1984.

Bloom, Howard S. "The Impact of Adult-Oriented Job-Training Programs: A Synthesis of the Best Available Evidence." Cambridge, MA: John F. Kennedy School of Government, Harvard University. December 1984.

Bloom, Howard S. "Estimating the Effect of Job-Training Programs, Using Longitudinal Data: Ashenfelter's Findings Reconsidered." Journal of Human Resources. Fall 1984.

Bloom, Howard S. and Maureen A. McLaughlin. CETA Training Programs--Do They Work for Adults? Washington, D.C.: Congressional Budget Office and National Commission for Employment Policy. July 1982.

Boruch, Robert F. "Solutions to Legal and Ethical Problems in Applied Social Research: Perspective and Prospects." in Boruch and Cecil.

Boruch, Robert F. "Equity and Randomized Experiments" in Joan E. Sieber, Editor. NIH Readings on the Protection of Human Subjects in Behavioral and Social Science Research. Frederick, MD: University Publications of America. 1984.

Boruch, Robert F. and Joe S. Cecil, Editors. Solutions to Ethical and Legal Problems in Social Research. New York: Academic Press. 1983.

Breger, Marshall J. "Randomized Social Experiments and the Law." in Boruch and Cecil.

Bryant, Edward C. and Kalman Rupp. "Evaluation of the Job Training Partnership Act: Implications of the CETA Experience." Rockville, MD: Westat. January 1985.

Deming, W. The Theory of Sampling. New York: Dover. 1950.

Dickinson, Katherine P., Terry R. Johnson and Richard W. West. "An Analysis of the Impact of CETA Programs on Participants' Earnings." Menlo Park, CA: SRI-International. November 1984.

Dickinson, Katherine P., Terry R. Johnson and Richard W. West. "Summary of an Analysis of the Impact of CETA Programs on Participants' Earnings and Implications for Evaluating JTPA." Menlo Park, CA: SRI-International. January 1985.

Finifter, David H. "Estimating Net Earnings Impact of Federally Subsidized Employment and Training Programs: What We Have Learned and Where Do We Go From Here?" Williamsburg, VA: Department of Economics. College of William and Mary. February 1983.

Fraker, Thomas and Rebecca Maynard. "The Use of Comparison Group Designs in Evaluations of Employment-Related Programs." Princeton, NJ: Mathematica Policy Research, Inc. January 1985.

Fraker, Thomas and Rebecca Maynard with Lyle Nelson. An Assessment of Alternative Comparison Group Methodologies for Evaluating Employment and Training Programs. Final Report. Princeton, NJ: Mathematica Policy Research, Inc. June 23, 1984.

Geraci, Vincent J. Short-Term Indicators of Job Training Program Effects on Long-Term Participant Earnings. Project Working Paper 2. Austin, TX: Center for Economic Research. August 1984.

Hausman, Jerry A. and David A. Wise. 1985. Editors. Social Experimentation. Chicago, IL: The University of Chicago Press.

Hausman, Jerry A. and David A. Wise. 1985. "Technical Problems in Social Experimentation: Cost versus Ease of Analysis." In Hausman and Wise. 1985.

Heckman, James J. and Richard Robb, Jr. "Alternative Methods for Evaluating the Impact of Training on Earnings." Chicago, IL: Department of Economics. University of Chicago. December 1983.

King, Christopher T. and Vincent J. Geraci. Short-Term Indicators of Job Training Program Performance: Policy Issues, Past Studies and Research Agenda." Austin, TX: Center for Economic Research.

LaLonde, Robert J. Evaluating the Econometric Evaluations of Training Programs with Experimental Data. Working Paper No. 183. Princeton, NJ: Industrial Relations Section. Princeton University. November 1984.

Maynard, Rebecca and Thomas Fraker. "Notes for Talk Presented at the JTLS Advisory Panel Meeting" Princeton, NJ: Mathematica Policy Research, Inc. January 15, 1985.

Pitcher, Hugh M. "A Sensitivity Analysis to Determine Sample Sizes for Performing Impact Evaluation of the CETA Programs." In Farrell Bloch, Editor. Evaluating Manpower Training Programs. Supplement 1. Greenwich, CT: JAI Press, Inc. 1979.

Rosenbaum, Paul R. and Donald B. Rubin. "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score." The American Statistician. Vol. 39. No. 1. February 1985.

Sedransk, J. "Designing Some Multifactor Analytic Studies." Journal of American Statistical Association. 62: pp. 1121-1139. 1967.

Simms, Margaret C. and Laurie J. Bassi. "Data Needs for Evaluating JTPA." Washington, D.C.: The Urban Institute. February 25, 1985.

Stafford, Frank P. 1985. "Income Maintenance Policy and Work Effort: Learning from Experiments and Labor Market Studies." In Hausman and Wise. 1985.

United States Congress. Job Training Partnership Act. Public Law 97-300. October 13, 1982.

United States Congress. U.S. Senate Committee on Post Office and Civil Service. Title 13, United States Code Census. Committee Print No. 2. 94th Congress. 2d. Session. December 31, 1976. Washington, D.C.: U.S. Government Printing Office, 1977.

Westat. Continuous Longitudinal Manpower Survey: Summary of Net Impact Results. Rockville, MD. April 1984.

Westat. Job Training Longitudinal Survey: Summary Description of JTLS Study Design. Rockville, MD. September 1984.